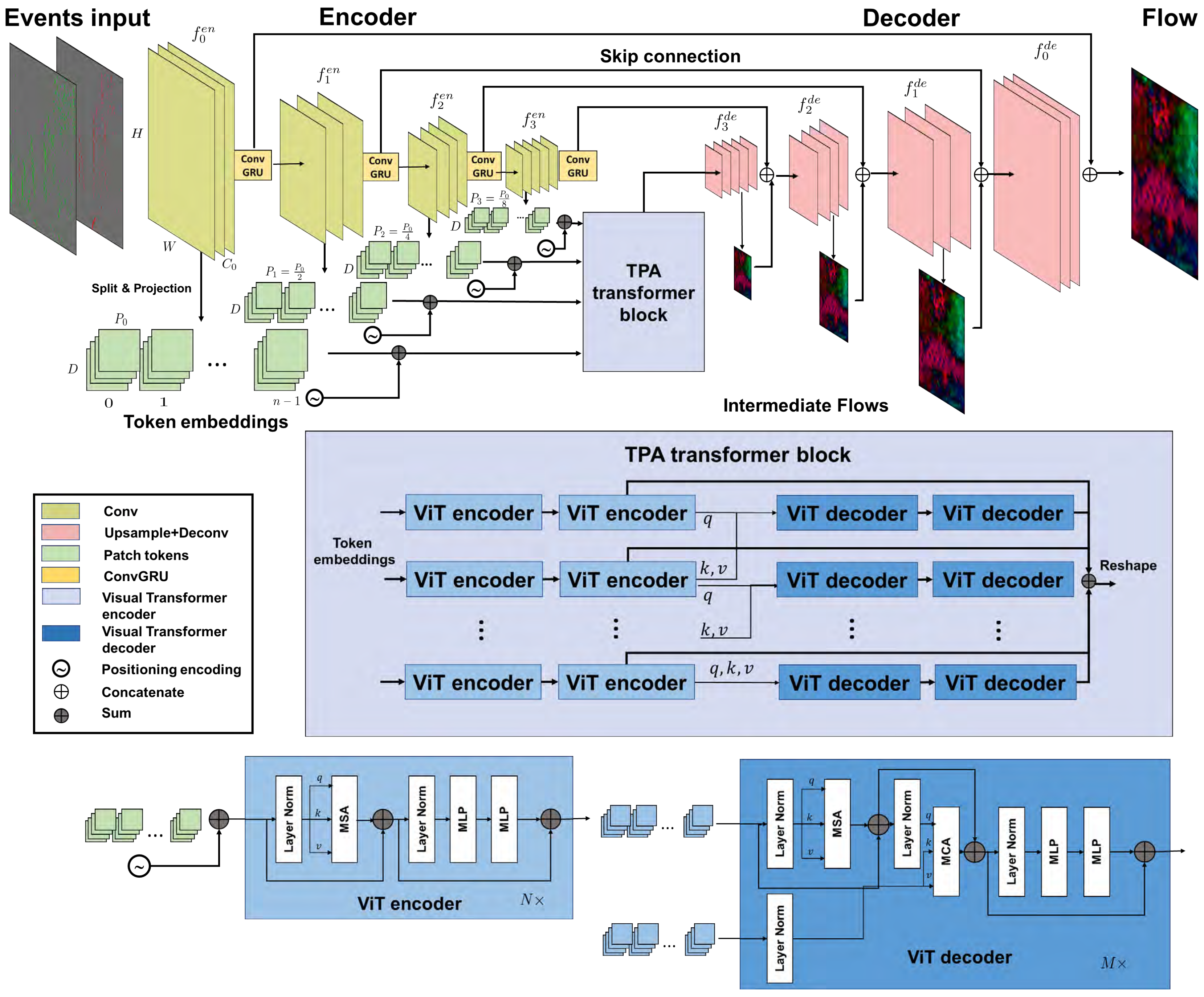
Event Transformer FlowNet for Optical Flow Estimation

Yi Tian, Juan Andrade-Cetto

1. Abstract

Event cameras are bioinspired sensors that produce asynchronous and sparse streams of events at image locations where intensity change is detected. They can detect fast motion with low latency, high dynamic range, and low power consumption. Over the past decade, efforts have been conducted in developing solutions with event cameras for robotics applications. In this work, we address their use for fast and robust computation of **optical flow**. We present **ET-FlowNet**, a hybrid RNN-ViT architecture for optical flow estimation. Visual transformers (ViTs) are ideal candidates for the learning of global context in visual tasks, and we argue that rigid body motion is a prime case for the use of ViTs since long-range dependencies in the image hold during rigid body motion. We perform end-to-end training with **self-supervised learning using contrast maximization loss**. Our results show comparable and in some cases exceeding performance with state-of-the-art event-based optical flow estimation methods.

2. Network Architecture



We follow a U-Net encoder-decoder architecture as in most of the literature, use **Conv-GRU units** to extract the temporal information and incorporate **a token pyramid aggregation (TPA) transformer block** to extract the global spatial context. The TPA module connects to all feature map outputs from the encoders. The transformer encoders model the internal spatial dependency of each feature map while the transformer decoders capture the interactions among all the feature maps of different scales. In such a way, we fully model the interactions across time, space and scales.

5. Ablation study

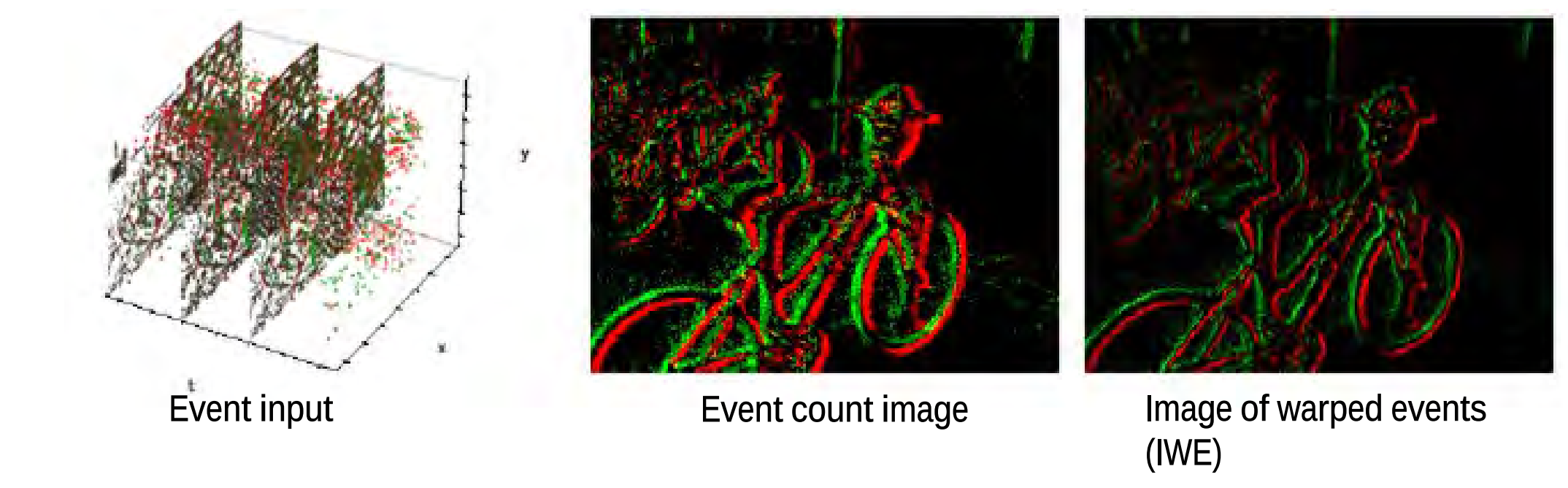
	outdoor_day1		indoor_flying1		indoor_flying2		indoor_flying3	
dt = 1 frame	AEE	% Outlier	AEE	% Outlier	AEE	% Outlier	AEE	% Outlier
baseline_2R	0.46	0.16	0.60	**0.40**	1.22	8.61	0.96	**5.58**
baseline_4R	0.82	1.21	0.88	1.50	1.39	9.88	1.16	6.64
ET-FlowNet_2T	0.41	0.14	0.59	0.50	**1.19**	8.81	0.99	6.79
ET-FlowNet_4T	**0.39**	**0.12**	**0.57**	0.53	1.2	**8.48**	**0.95**	5.73
ET-FlowNet_trapezoid	0.51	0.26	0.81	1.96	1.59	14.97	1.31	11.51
dt = 4 frame	AEE	% Outlier	AEE	% Outlier	AEE	% Outlier	AEE	% Outlier
baseline_2R	1.68	11.77	2.26	23.88	4.08	44.60	3.20	33.98
baseline_4R	2.97	39.41	3.51	47.73	4.92	60.24	4.20	53.00
ET-FlowNet_2T	1.55	10.67	2.17	22.61	4.00	42.82	3.31	35.27
ET-FlowNet_4T	**1.47**	**9.17**	**2.08**	**20.02**	**3.99**	**41.33**	**3.13**	**31.70**
ET-FlowNet_trapezoid	1.87	16.15	3.02	35.44	5.51	55.20	4.46	47.30

Ablation studies for ET-FlowNet with quantitative evaluation on the MVSEC dataset for dt = 1 and dt = 4. 2R and 4R stand for two or four residual blocks, while 2T and 4T stand for two or four transformer blocks, respectively. Our model with the best performance (ET-FlowNet_4T) is marked in red box.

Further, we perform ablation studies based on three factors: a) **with or without the transformer block** b) **the number of encoders and decoders** included in the transformer block, and c) **their stacking structure**. The first factor is to show the effectiveness of the transformer block. We compare ET-FlowNet with the best-performing self-supervised learning model to date for different training/testing sequences, ConvGRU-EV-FlowNet. We retrain the ConvGRU-EV-FlowNet under our settings as the baseline model. Our model (ET-FlowNet_4T) beats the baseline models in most of the sequences for the dt = 1 case and in all the sequences for the dt = 4 case.

The other two factors are aimed at identifying the proper design of the TPA transformer block. In our case, the variant with 4 transformer blocks (2 encoders and 2 decoders) outperforms an architecture with only 2 transformer blocks (1 encoder and 1 decoder) by a small margin in most sequences. For thestacking structure, we compare a square stacking structure (4-4-4-4) with 2 ViT encoders and 2 ViT decoders for each pyramid scale, with a trapezoid stacking one (6-4-4-2) with 3 ViT encoders and 3 ViT decoders in the last scale and 1 for each in the first scale. The conclusion is that a stacking structure is preferred over the trapezoid one.

3. Training method – self supervised learning with Contrast Maximization loss



Average timestamp

$$T_{p'}(x, y \mid t') = \frac{\sum_{i|p_i=p'} \kappa(x - x'_i)\kappa(y - y'_i)t_i}{\sum_{i|p_i=p'} \kappa(x - x'_i)\kappa(y - y'_i) + \epsilon} \qquad p' \in \{+,-\}, \epsilon \approx 0 \qquad (1)$$

Contrast Maximization loss

$$\mathcal{L}_{CMax}(t') = \frac{\sum_x \sum_y T_{p'\in\{+\}}(x, y \mid t')^2 + T_{p'\in\{-\}}(x, y \mid t')^2}{\sum_x \sum_y [n(x') > 0] + \epsilon} \qquad \epsilon \approx 0\ . \qquad (2)$$

Flow loss

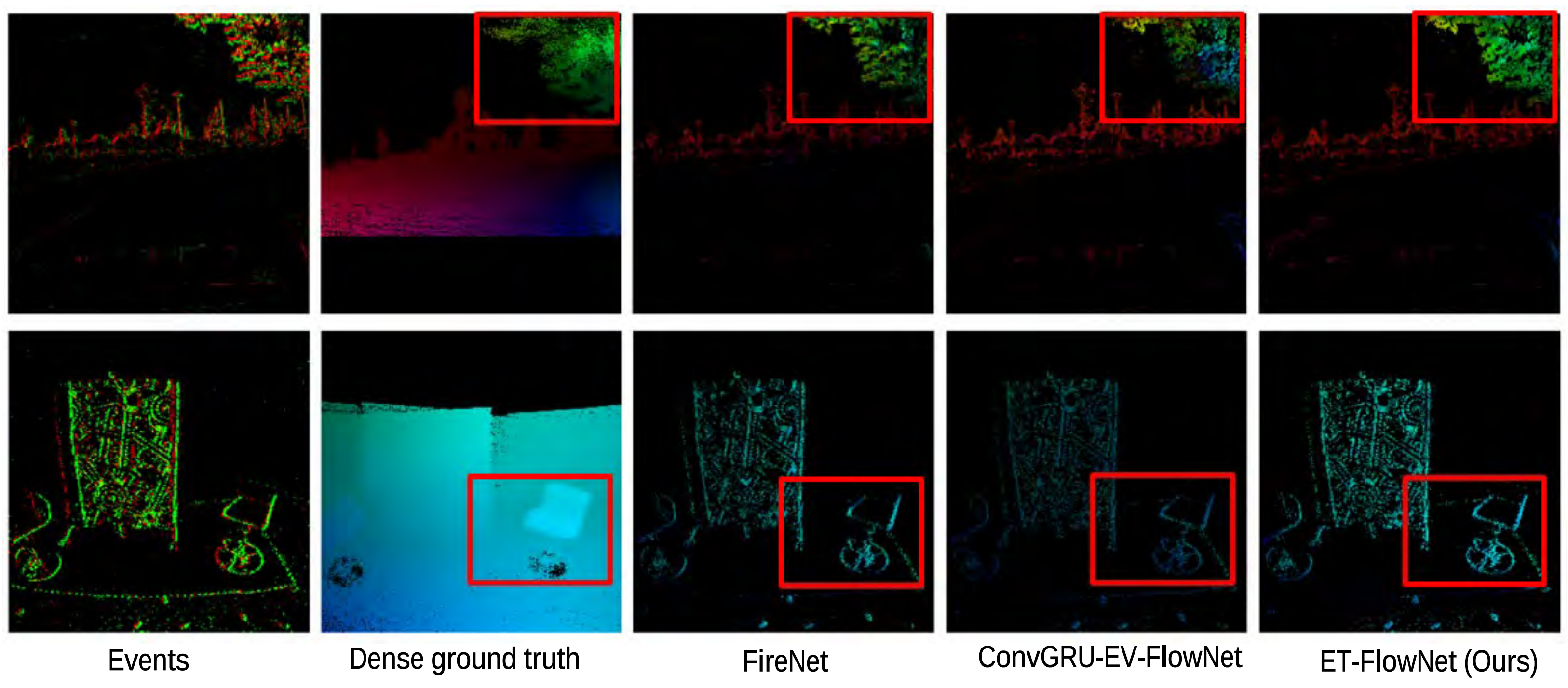
$$\mathcal{L}_{flow} = \mathcal{L}_{CMax}^{forward}(t') + \mathcal{L}_{CMax}^{backward}(t') + \lambda\mathcal{L}_{smooth}\ . \qquad (3)$$

We train the network using self supervised learning with **Contrast Maximization loss**. The method aims to find the optimized parameters that compensate for the motion and provide a **deblurred image of warped events (IWE)** upon convergence. We first generate the per-pixel and per-polarity average timestamp of the IWE via bilinear interpolation. The contrast maximization loss is expressed as the scaled sum of the squared temporal images both in the forward and backward direction. In addition, we add a smooth term in the neighborhood pixels as regularizer.

4. Qualitative and quantitative results

		outdoor_day1		indoor_flying1		indoor_flying2		indoor_flying3		
Training	$dt = 1$ frame	AEE	% Outlier	AEE	% Outlier	AEE	% Outlier	AEE	% Outlier	Learning
MVSEC	EV-FlowNet [35]	0.49	0.20	1.03	2.20	1.72	15.10	1.53	11.90	Semi-SL
	Spike-FlowNet [16]	0.47	-	0.84	-	1.28	-	1.11	-	
	STE-FlowNet [8]	0.42	0.00	0.57	0.1	0.79	1.6	0.72	1.3	
	EV-FlowNet2 [37]	**0.32**	**0.00**	0.58	**0.00**	**1.02**	**4.00**	**0.87**	**3.00**	Self-SL
	EV-FlowNet2_indoor [24]	0.36	0.09	-	-	-	-	-	-	
FPV	EV-FlowNet2_retrained	0.56	0.17	0.62	**0.26**	**1.10**	**5.97**	**0.90**	**3.54**	
	ConvGRU-EV-FlowNet [12]	0.47	0.25	0.60	0.51	1.17	8.06	0.93	5.64	
	FireNet [12]	0.55	0.35	0.89	1.93	1.62	14.65	1.35	10.64	
	ET-FlowNet (ours)	**0.39**	**0.12**	**0.57**	0.53	1.2	8.48	0.95	5.73	
	$dt = 4$ frames	AEE	% Outlier	AEE	% Outlier	AEE	% Outlier	AEE	% Outlier	
MVSEC	EV-FlowNet [35]	1.23	7.30	2.25	24.70	4.05	45.30	3.45	39.70	Semi-SL
	Spike-FlowNet [16]	1.09	-	2.24	-	3.83	-	3.18	-	
	STE-FlowNet [8]	0.99	3.9	1.77	14.7	2.52	26.1	2.23	22.1	
	EV-FlowNet2 [37]	**1.30**	9.70	2.18	24.20	**3.85**	46.80	3.18	47.80	Self-SL
	EV-FlowNet2_indoor [24]	1.49	11.72	-	-	-	-	-	-	
FPV	EV-FlowNet2_retrained	2.14	20.76	2.35	26.35	3.92	47.84	3.18	37.47	
	ConvGRU-EV-FlowNet [12]	1.69	12.50	2.16	21.51	**3.90**	**40.72**	**3.00**	**29.60**	
	FireNet [12]	2.04	20.93	3.35	42.5	5.71	61.03	4.68	53.42	
	ET-FlowNet (ours)	**1.47**	**9.17**	**2.08**	**20.02**	3.99	41.33	3.13	31.70	

Quantitative results for optical flow estimation for dt = 1 and dt = 4, and tested on various sequences of the MVSEC dataset. We sorted the methods according to: a) the training dataset used: MVSEC or UZH-FPV; and b) the learning method: semi-supervised with grayscale images or self-supervised using event data only. Semi-supervised results are only shown for baseline purposes. Our method is compared with other self-supervised methods, with the best performing one shown in black bold. We highlight in blue methods matching the same learning and training conditions as ours (self-supervised and trained on the UZH-FPV dataset). Of these, the best-performing method is highlighted in blue bold unless they are already highlighted in black bold. To relate the success of self-supervised methods to semi-supervised ones, we also underline results for the overall winners in each tested sequence.



Qualitative results for optical flow evaluated on the MVSEC dataset for the dt = 1 case. The first row is from outdoor_day1 sequence and the last row is from the indoor_flying sequence. Note that for evaluation we use the masked sparse optical flow.

We train our model on the **FPV dataset**, following the same pipeline as in ConvGRU-EV-FlowNet and FireNet. Thus, these two networks are our main target for evaluation comparison. In general, our method outperforms FireNet in all sequences and beats ConvGRU-EV-FlowNet, especially with a large margin for the outdoor_day1 sequence, and yields similar results in the indoor sequences.

The qualitative results on the sequences further confirm these numbers with a visual comparison with ground-truth data Our results show more consistency compared to ground truth, especially these sparse and far away objects thanks to the transformer block that capture better the spatial dependency.

6. Conclusion

In this paper, we proposed ET-FlowNet, the first RNN-ViT framework for event-based optical flow estimation. Our network incorporates a ViT block with TPA to extract the global spatial context and interaction among the multi-scale outputs from the encoder. We perform qualitative and quantitative evaluations on various datasets and compare them to state-of-the-art methods. Our method achieves superior results to other self-supervised methods on some of the sequences when trained and tested on different datasets. In addition, we use the event count representation to simplify the event preprocessing step and to provide a fair comparison with prior work. In future work, we aim to simplify the complexity of the network for less memory consumption and expand the self-attention mechanism as the backbone for the temporal domain.

Date: November 21-24, 2022

This work was supported by projects EBSLAM DPI2017-89564-P and EBCON PID2020-119244GB-I00 funded by MCIN/AEI/10.13039/501100011033 and by an FI AGAUR PhD grant to Yi Tian.